DOCUMENT RESUME

ED 191 870 TM 800 485

AUTHOR Divgi, D. R.
TITLE Evaluation of Scales for Standardized Tests.
PUB DATE Apr 80
NOTE 18p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Equated Scores; *Evaluation Criteria; Item Analysis; Latent Trait Theory; *Mathematical Models; *Scoring Formulas; *Standardized Tests; Test Interpretation; Test Theory
IDENTIFIERS Out of Level Testing

ABSTRACT

A method is proposed for providing an absolute, in contrast to comparative, evaluation of how well two tests are equated by transforming their raw scores into a particular common scale. The method is direct, not requiring creation of a standard for comparison; expresses its results in scaled rather than raw scores, and allows examination of the scale at as many different abilities as required, rather than calculating a single overall statistic. The method was tried on two levels of a vocabulary test. A latent trait model was used, the levels having some items in common. These items were used to place parameters of the two levels on a unified ability scale. By considering different values of ability, a detailed description of how well the tests had been equated could be obtained. Reading scores were used as predictors, cases to be used were selected at random, and the LOGIST program was used for estimation of parameters. Results illustrated how floor and ceiling effects result in higher scaled scores for those who take the higher level. (Author/GK)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED191870

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

# Evaluation of Scales for Standardized Tests

D. R. Divgi

Syracuse University

and

The University of Iowa

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. R. Divgi

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 804485

Revised version of a paper presented at the 1980 Annual Meeting of the American Educational Research Association.

Abstract

In any multilevel test battery, raw scores on different levels are equated and then transformed into scores on a common scale. Ideally it should not matter which of two levels an examinee takes. How far reality differs from the ideal can be judged by using a latent trait model, provided the levels have some items in common. The common items are used to place parameters of the two levels on a unified ability scale. Then, at any given ability, one calculates expected scaled scores from the two levels. Any difference between the expected values implies that, for persons of that ability, the two levels are not equivalent and therefore the scale is biased. By considering many different values of ability, one obtains a detailed description of how well the tests have been equated. This procedure was applied to a vocabulary test. The results illustrate how floor and ceiling effects result in higher scaled scores for those who take the higher level.

# Evaluation of Scales for Multilevel Test Batteries

Equating and scaling are essential in standardized testing. If students are to be tested in different grades and their progress evaluated, the test battery must contain different levels for different grades. Higher levels are more difficult than lower ones. Scores at different levels can be compared only if formulas are available for converting these scores into one another or, more conveniently, into a common scale spanning all the levels. Scores at different levels are converted into the same scaled score if they are considered to be equivalent according to some equating method.

While many methods are available for equating and scaling (Angoff, 1971; Lord, 1977), there are few criteria for evaluating the results. Slinde and Linn (1977) picked out pairs of raw scores that were equivalent according to the Anchor Test Study (Bianchini & Loret, 1974), looked up the corresponding scaled scores provided by the test publishers, and found the latter to differ considerably in many cases. Whereas Slinde and Linn concluded that these discrepancies indicated a fault in the anchor test method, it is quite possible that the defects lie in the publishers' scales. A major study of anchor test methods has been carried out by Marco, Petersen & Stewart (1979). They judged the results of an equating procedure by comparing them with results from two "criterion equatings", one of which used item response theory while the other did not. Both equatings were based on the equipercentile procedure with test-retest, and the same large sample; yet their results differed noticeably. If the criterion equatings were identical, both would yield the same discrepancy index for any given method. In fact the indices were ~~noticeably~~ appreciably different, thus demonstrating the basic drawback of comparisons with "criterion" equatings. Since no single equating procedure is generally accepted as the best, the results of any comparative evaluation depend on the subjective

choice of a criterion method. Even if there were a consensus about the preferred procedure, a basic question would still remain: Is the best method good enough? ("The results obtained by conventional methods cannot be justified as a criterion" - Lord, 1977, p. 132.)

## An Index for Evaluating a Scale

It is possible to evaluate an equating procedure on its own, by comparing it with the ideal rather than with another procedure. Ideal equating has been defined in the literature. According to Angoff (1971, p. 563), "Two scores, one on Form X and the other on Form Y ..., may be considered equivalent if their corresponding percentile ranks in any given group are equal." Lord (1977, p. 128) states the idea in stronger terms: "Transformed scores $y^*$ and raw scores $x$ can be called 'equated' if and only if it is a matter of indifference to each examinee whether he is to take test X or test Y. ... A necessary but not sufficient condition for $y^*$ and $x$ to be'equivalent' scores, (i.e., to be equated) is that $y^*$ and $x$ have identical frequency distributions in all groups and subgroups of interest." Equating in this sense is impossible if tests differ in reliability (Angoff, 1971, p. 568). If they differ in difficulty, their reliabilities will not be equal in all subgroups. The easy test is better at low abilities and the difficult test at high abilities. Therefore, although one can determine the relationship between true scores, observed scores cannot be equated (Lord, 1977, p. 128).

Given that the results of any equating procedure will be imperfect, we can evaluate them by quantifying the difference between the actual and the ideal. The index to be used should be easy to interpret. A good candidate is the difference between means of $y^*$ and $x$. However, one must remember that differences between means are affected by nonlinear transformations. What is reported to the persons concerned, and what is generally used to make decisions, is not $y^*$ or $x$ but the score on a common scale into which both raw scores are transformed. When X and Y are two different levels of a test battery, ~~with a~~

differing substantially in difficulty, the transformation is nonlinear. Different transformations yield different scales such as standard scores and grade equivalent (GE) scores, which differ not only in mathematical properties but even in meaning (Echternacht, 1977; Gardner, 1962, p. 13). The results of Slinde and Linn (1977) clearly show the effects of these differences. In Column 2 of their Table 1, scaled score differences of 7 and 4 equal GE differences of 0.0 and 1.2 respectively! Therefore any evaluation ought to be based on scores on the common scale and not on raw scores. Moreover, if two scales are provided for the same battery, each should be evaluated separately.

Let $S_X(\underline{x})$ be the transformation of raw scores on Level X into the common scale. Let $\underline{y}^*(\underline{y})$ be the score on X which is equated to raw score $\underline{y}$ on Y. Then the transformation of $\underline{y}$ to the common scale is $S_Y(\underline{y}) = S_X[\underline{y}^*(y)]$. Assume that ~~distributions~~ probabilities $P_X(\underline{x})$ and $P_Y(\underline{y})$ are available in a population or subpopulation such as a school district. The mean error due to imperfect equating for this population is

Equation 1 here

The difference represents equating bias for the subpopulation. If $\underline{D}$ is positive, the group as a whole will look better on test X than on test Y; in other words the scale is biased in favor of those who take test X. The ~~phrase "equating bias" is~~ word "bias" is used rather than "equating error" because the latter emphasizes random error at the time of equating (Angoff, 1971) while the present emphasis is on systematic effects at the time of use.

In order to determine bias in a scale using empirical proportions, each examinee must take both tests, or the tests must be given to random samples from the same population. Such data are rarely available when vertical equating is involved, because different levels are used in different grades. However, one can obtain the necessary distributions theoretically. Then, at any ability $\theta$, one can calculate

Equation 2 here

Thus bias can be calculated as an explicit function of ability, in contrast to the use of empirical distributions which yield only an average over an unknown distribution of ability. This is an important advantage of calculating bias theoretically.

## Calculation of $D(\theta)$

The distributions of raw scores $x$ and $y$ can be calculated using one of the latent trait models. The most promising of these is the three parameter logistic model (Birnbaum, 1968). Item g is described by the discrimination parameter $a_g$, difficulty $b_g$ and guessing parameter $c_g$. The probability that a person with ability $\theta$ will answer this item correctly is

Equation 3 here

where $g = 1, 2, \ldots, n$ and $\psi$ is the logistic function

Equation 4 here

$\theta$ can have any real value. (The factor $D = 1.7$ used by Birnbaum is omitted for simplicity. It is needed only while comparing logistic and normal ogive models.)

Once values of $P_g(\theta)$ are available for all items, the distribution of observed scores can be calculated. Exact calculation would be expensive in terms of computer time, but it is not necessary. The compound binomial approximation (Lord, 1965) should be satisfactory. The mean probability of a correct answer (which is the proportion-correct true score) equals

Equation 5 here

Let

Equation 6 here

and

Equation 7 here

Then, using Lord's (1965) Equations (19), (22) and (26), the compound binomial

probability is

Equation 8 here

where

Equation 9 here

(Lord's assumption that $V_2 \propto \zeta(1-\zeta)$ is unnecessary here because a much more detailed model is being used.) Once probabilities $P_X(\underline{x}|\theta)$ and $P_Y(\underline{y}|\theta)$ are available for raw scores, one can calculate expected scaled scores on the two levels. The difference $\underline{D}(\theta)$ thus obtained indicates how effectively the scale equates the levels for examinees with ability $\theta$. If $\underline{D}(\theta)$ is positive, persons with ability $\theta$ will tend to get higher scaled scores on level X than on level Y.

The probabilities can be used to calculate error variances also. A more useful concept is "information" (Birnbaum, 1968). This is a measure of how well a score reveals differences in ability. Birnbaum has shown how to calculate the information function of a linear combination of item scores. A scaled score, involving a nonlinear transformation of the number-right score, does not permit use of Birnbaum's formula. However, one can use the basic definition

Equation 10 here

where the denominator is the error variance of the scaled score. The numerator can be computed using

Equation 11 here

Thus latent trait theory provides two measures for evaluating a scale - its bias (if any) and the information in the scores obtained from the two levels of the test.

## Practical Considerations

In the previous section it was assumed that item parameters were available. It is not a trivial matter to obtain parameter estimates which are suitable for comparing different levels. First one must obtain estimates for each level separately, using large samples (preferably around 2000). One method of doing

this is to use the LOGIST program (Wood, Wingersky & Lord, 1976). These estimates are arbitrary to some extent because the value of $P_g(\theta)$ remains unchanged by transformations of the following form:

Equations 12a, 12b, 12c here

Therefore all parameter sets related by transformations of this type are equivalent. Thus the scale and origin of ability are arbitrary. One can equalize these for two parameter sets if they have some items in common. These common ("anchor") items may be parts of the tests themselves, or external items added for equating purposes.

The common items get calibrated twice. Since the two sets of values estimate the same parameters, they are used to determine $\alpha$ and $\beta$. This can be done in different ways. Marco (1977, p. 141) used the standard deviations of difficulty estimates to calculate $\alpha$. One may as well use the ratio of the (arithmetic or geometric) means of discrimination values, or fit a line through the scatterplot of difficulty estimates. These methods yield somewhat different values, and there are no a priori grounds for selecting one method as the best. Moreover, all these methods are sensitive to bad items. Sometimes the discrimination parameter is so small that the item characteristic curve (ICC) looks like a straight line over the ability range of interest, and the difficulty estimates are very large or very small. (See Lord, 1977, Figures 8 and 9, for examples.) One such item can change $\alpha$ and $\beta$ by a large amount. Therefore Haebara (1980) has suggested that, instead of trying to equate two sets of estimates, one should look at the resulting ICCs themselves. For each common item, parameter estimates from the X calibration yield one ICC; those from the Y calibration, transformed as in Equations (12a) and (12b), yield another. If estimates were error-free, the two curves would coincide after proper choice of $\alpha$ and $\beta$. Haebara averages the squared difference over a suitable distribution of $\theta$, sums over items, and then fits $\alpha$ and $\beta$ by minimizing this sum. Because

Haebara's method involves an arbitrarily chosen ability distribution, a minimax modification was used in this study. Instead of using individual ICCs, they were added to obtain expected total score as a function of $\theta$. The two calibrations (X and transformed Y) provided two characteristic curves for the anchor test. Values of $\alpha$ and $\beta$ were selected to minimize the largest distance between these curves. This maximum distance cannot be less that that between chance scores. The latter is due entirely to random differences between estimates of guessing parameters, and has nothing to do with scale or origin of ability. This extraneous complication was removed by averaging the two estimates of $c_g$ for each common item (only while fitting $\alpha$ and $\beta$). In summary, ability scales of the X and Y calibrations were equated by minimizing the maximum difference between expected anchor test scores based on item parameter estimates $\left[\hat{a}_{gX}, \hat{b}_{gX}, (\hat{c}_{gX} + \hat{c}_{gY})/2\right]$ and $\left[\hat{a}_{gY}/\alpha, (\hat{b}_{gY} + \beta), (\hat{c}_{gX} + \hat{c}_{gY})/2\right]$.

Another problem is that of dealing with omitted and unreached items. LOGIST ignores the items an examinee did not reach, and to each omitted item assigns a score of 1/A where A is the number of options (Wood et al, 1976, p. 9). This is justified if the items are to be used in future under conditions where every examinee will answer every question (tailored testing, for example). However, this procedure is not appropriate if the item parameters will be used to predict score distributions and other features of a test on which the score is defined as the number of correct answers (i.e., blanks are treated as wrong answers). As an extreme example, suppose the last item is attempted by only half the examinees while the first item is answered by all. Assume that, on both items, half the answers are correct. Then LOGIST will consider the two items to be equally difficult. There is nothing in the Birnbaum model to predict whether an examinee of ability $\theta$ will reach and attempt a particular question. (The test is assumed to be unspeeded.) In theoretical calculations one has no choice but to assume that all items are answered. Then the parameter estimates from

LOGIST will predict equal numbers of correct answers on the first and last items while in reality the ratio will be two to one. Therefore, if the model will be used to calculate distributions of number-right scores, all items not answered correctly should be treated as wrong. (Marco et al, using real data, calculated estimated true score by summing $P_g(\theta)$ over attempted items only. This is not possible in theoretical calculations.)

There is another consideration, however. If the test is constructed with the intention of using the common items for equating, they should be placed at random throughout each level. If, on the other hand, use of these items is an afterthought, their locations in the two levels may differ systematically. In the tests used in this study, the common items are at the beginning of the higher level (X) and at the end of the lower level (Y). Therefore, if both tests are given to the same sample, p values of the common items will tend to be slightly higher in the X data than in the Y data, because some examinees do not reach items near the end. Smaller p values lead to higher estimates of b's. This is a source of systematic error, making $\beta$ smaller than it should be. This affects all subsequent results. Such systematic error should be avoided by leaving unreached items out of the analysis.

To summarize, two issues are involved. (a) The task of a model is to predict what will happen in practice. Therefore if unanswered questions are scored as wrong in actual use, the same ought to be done while calibrating the model. (b) If the locations of anchor items in the two tests differ systematically, and some examinees fail to reach all items, this is a source of systematic error in $\beta$. The error can be avoided by ignoring unreached items. If locations of anchor items in the test are random, argument (b) does not arise. If the locations differ systematically, (b) should override (a).

## Illustration

The procedure recommended above was tried on two levels of a vocabulary test, X being the higher level. The numbers of items were $n_X = 44$ and $n_Y = 42$, with 17 items present in both levels. Available data sets contained 19,640 and 17,366 usable cases respectively. Samples of about 2,000 each were drawn. For given sample size, parameter estimates are more accurate if examinees with very high and very low abilities are present in sizable proportions (Lord, 1968, p. 1017). Therefore Reading scores were used as predictors. Cases to be used were selected at random, with probability of selection depending on the Reading score in such a way that the distribution of Reading scores was almost flat in each sample. Sample statistics for X and Y were as follows: Sample sizes 1980 and 1978; mean scores 26.3 and 24.3; standard deviations 11.0 and 11.6; KR-20 reliabilities .94 and .95 respectively. All items after the last answered one were treated as unreached, and other unanswered items were scored as wrong. The LOGIST program was used for estimation of parameters. The 17 common items were used to compare the scales and origins of the two calibrations. Parameters for level Y were transformed as in Equations (12a) and (12b) with $\alpha = 0.974$ and $\beta = -0.705$. The scaled score used was the grade equivalent (GE) score. For values of $\theta$ from -4 to 4, expected values, error variances, and information functions were calculated for both raw (number-right) and GE scores.

Results are shown in Table 1. All values of $D(\theta)$ turned out to be positive; i.e., expected GE scores are higher for those who take the higher level. Major features are these:

1. $D(\theta)$ is large at very small values of $\theta$. Chance scores are about equal on the two levels, test difficulty having little effect. However a given raw score corresponds to higher GE on the higher level. The precise values of $D(\theta)$ are not important at very low abilities. These are determined almost entirely by

guessing parameters, and estimates of these are not accurate (Wood et al, 1976, p. 5). In the present case, values of $c_g$ were larger by about .05 in the higher level.

2. $D(\theta)$ is large again at very high abilities, approaching the asymptotic value of 27 which is the difference between GE values corresponding to perfect scores on the two levels. At any given ability, the higher level yields a lower expected (proportion-correct) raw score. However a raw score on the higher level corresponds to higher GE than the same score on the lower level. These two facts tend to counterbalance each other. However, as ability becomes very high, both expected raw scores approach the ceiling, and the additional difficulty of the higher level has very little effect.

3. These opposing tendencies almost cancel each other around $\theta = -1$. Thus best equating is achieved at middle ability levels.

4. GE scores contain slightly less information than raw scores. The reason may be that each GE score is rounded to an integer and therefore the relation of GE to raw score is not smooth. Otherwise the transformation from raw to GE scores does not appear to affect information content.

The qualitative features of the results are not specific to the tests used in this study. They follow from the nature of multiple-choice tests with different levels of difficulty, and similar results will be obtained with any other test battery. Only the actual numerical values will differ from one pair of tests to another.

## Summary

The method proposed here provides an absolute (in contrast to comparative) evaluation of how well two tests are equated by transforming their raw scores into a particular common scale. It has the following useful features.

1. The method is direct; it does not require creation of a standard for comparison, such as the criterion equatings used by Marco et al (1979).

2. Results are expressed in terms of scaled scores rather than raw scores. The former are considered easier to interpret, and they are the ones for which norms are provided by publishers.

3. Instead of calculating a single over-all statistic, one can examine the scale at as many different abilities as required. This makes it possible to show clearly that equating is good at medium ability levels but not at high and low levels. Such detailed information is useful if the test battery is to be used with groups of different ability levels.

When one test is more difficult than the other, floor and ceiling effects make it inevitable that persons with very low and very high abilities will tend to obtain higher scaled scores if they take the more difficult test. Because of this bias, one must be cautious in interpreting results that involve different levels of a standardized test battery. (Similar reservations are expressed by Slinde and Linn, 1977.) It is always possible that part of the observed improvement in performance resulted from using a higher level in the posttest than in the pretest.

Computer programs used in the study, written in FORTRAN for an IBM 370, are available from the author.

Acknowledgements

I am grateful to Prof. William Coffman and Prof. Melvin Novick for illuminating discussions.

References

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Bianchini, J. C. & Loret, P. G. Anchor test study. Final Report, 1974. (ERIC Documents Reproduction Service Numbers ED 092601 through 092634)

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Echternacht, G. Grade equivalent scores. Measurement in Education, 1977, 8(2), 1-4.

Gardner, E. F. Normative standard scores. Educational and Psychological Measurement, 1962, 22, 7-14.

Haebara, T. Equating logistic ability scales by a weighted least squares method. ITP Occasional Paper No. 27. The University of Iowa, Iowa Testing Programs, 1980.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Marco, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14, 139-160.

Marco, G. L., Petersen, N. S. & Stewart, E. E. A test of the adequacy of curvilinear score equating models. Paper presented at the 1979 Computerized Adaptive Testing Conference, Minneapolis, June 1979.

Slinde, J. A. & Linn, R. L. Vertically equated tests: Fact or phantom? Journal of Educational Measurement, 1977, 14, 23-32.

Wood, R L, Wingersky, M. S. & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. (RM-76-6). Princeton, N. J.: Educational Testing Service, 1976.

Table 1
Properties of Raw and GE Scores at Different Abilities[1]

| Ability θ | Bias in GE D(θ) | Raw Scores Mean | Raw Scores S.D. | Raw Scores Info | Grade Equivalents Mean | Grade Equivalents S.D. | Grade Equivalents Info |
|---|---|---|---|---|---|---|---|
| -4.00 | 10.3 | 9.7 | 2.7 | 0.0 | 44.7 | 9.4 | 0.0 |
| | | 8.6 | 2.6 | 0.0 | 34.4 | 7.3 | 0.0 |
| -2.00 | 5.5 | 10.9 | 2.8 | 0.5 | 48.7 | 9.5 | 0.5 |
| | | 12.0 | 2.9 | 3.6 | 43.2 | 6.7 | 3.5 |
| -1.75 | 4.0 | 11.6 | 2.9 | 1.1 | 50.7 | 9.4 | 1.1 |
| | | 13.6 | 3.0 | 6.7 | 46.7 | 6.1 | 6.5 |
| -1.50 | 2.7 | 12.5 | 2.9 | 2.1 | 53.6 | 9.2 | 2.1 |
| | | 15.9 | 3.1 | 11.3 | 50.9 | 5.3 | 11.0 |
| -1.25 | 2.1 | 13.7 | 3.0 | 3.8 | 57.5 | 8.8 | 3.8 |
| | | 18.7 | 3.1 | 16.8 | 55.4 | 4.5 | 16.5 |
| -1.00 | 2.2 | 15.4 | 3.0 | 6.4 | 62.2 | 8.1 | 6.3 |
| | | 22.2 | 3.1 | 22.0 | 59.9 | 3.9 | 21.8 |
| -0.75 | 3.0 | 17.6 | 3.1 | 10.1 | 67.5 | 7.0 | 9.9 |
| | | 25.9 | 3.0 | 25.2 | 64.5 | 3.7 | 25.0 |
| -0.50 | 4.0 | 20.3 | 3.1 | 14.9 | 73.2 | 6.1 | 14.8 |
| | | 29.5 | 2.8 | 25.3 | 69.2 | 3.7 | 24.9 |
| -0.25 | 5.5 | 23.5 | 3.1 | 20.2 | 79.3 | 5.7 | 20.1 |
| | | 32.8 | 2.5 | 22.2 | 73.8 | 4.2 | 21.1 |
| 0.00 | 7.2 | 27.1 | 3.0 | 24.5 | 86.0 | 5.5 | 24.5 |
| | | 35.4 | 2.2 | 17.0 | 78.8 | 5.2 | 15.7 |
| 0.25 | 8.8 | 30.7 | 2.8 | 26.2 | 92.8 | 5.4 | 26.0 |
| | | 37.4 | 1.9 | 12.0 | 84.0 | 6.1 | 11.2 |
| 0.50 | 10.8 | 34.1 | 2.6 | 23.9 | 99.7 | 5.7 | 23.5 |
| | | 38.8 | 1.6 | 8.1 | 88.9 | 6.4 | 7.7 |
| 0.75 | 13.6 | 36.9 | 2.3 | 18.9 | 106.5 | 6.0 | 18.6 |
| | | 39.7 | 1.4 | 5.4 | 92.9 | 6.2 | 5.2 |
| 1.00 | 16.4 | 39.0 | 1.9 | 13.6 | 112.4 | 5.9 | 13.4 |
| | | 40.4 | 1.2 | 3.6 | 96.0 | 5.8 | 3.5 |
| 1.25 | 18.9 | 40.6 | 1.7 | 9.5 | 117.2 | 5.6 | 9.2 |
| | | 40.9 | 1.0 | 2.3 | 98.3 | 5.3 | 2.3 |
| 1.50 | 20.9 | 41.6 | 1.4 | 6.5 | 121.0 | 5.4 | 6.3 |
| | | 41.2 | 0.9 | 1.6 | 100.1 | 4.8 | 1.5 |
| 2.00 | 24.0 | 42.9 | 1.0 | 3.0 | 126.3 | 4.6 | 2.9 |
| | | 41.6 | 0.6 | 0.7 | 102.3 | 3.9 | 0.7 |
| 4.00 | 26.9 | 43.9 | 0.2 | 0.1 | 131.6 | 1.4 | 0.1 |
| | | 42.0 | 0.2 | 0.0 | 104.7 | 1.6 | 0.0 |

[1]At each ability, first and second lines refer to higher and lower levels respectively. The bias equals expected GE score on the higher level minus that on the lower level.

$$D = \sum_x P_X(x)S_X(x) - \sum_y P_Y(y)S_Y(y). \qquad [1]$$

$$D(\theta) = E[S_X(x)|\theta] - E[S_Y(y)|\theta]. \qquad [2]$$

$$P_g(\theta) = c_g + (1-c_g)\,\psi[a_g(\theta-b_g)] \qquad [3]$$

$$\psi(z) = 1/[1 + \exp(-z)]\ . \qquad [4]$$

$$\zeta = (1/n) \sum_g P_g(\theta)\ . \qquad [5]$$

$$V_2 = (1/n) \sum_g [P_g(\theta) - \zeta]^2 \qquad [6]$$

$$p_n(x) = \binom{n}{x} \zeta^x(1-\zeta)^{n-x}\ . \qquad [7]$$

$$P(x|\theta) \simeq p_n(x)\ [1 + V_2C_2(x)] \qquad [8]$$

$$C_2(x) = [2(n-1)\zeta^2(1-\zeta)^2]^{-1}\ [-n(n-1)\zeta^2 + 2x(n-1)\zeta - x(x-1)]. \qquad [9]$$

$$I(\theta,S_X) = [dE(S_X|\theta)/d\theta]^2/\mathrm{var}(S_X|\theta) \qquad [10]$$

$$dE(S_X|\theta)/d\theta = \sum_x S_X(x)\ dP(x|\theta)/d\theta\ . \qquad [11]$$

$$a_g \rightarrow \alpha a_g \qquad \alpha>0\ , \qquad [12a]$$

$$b_g \rightarrow (b_g + \beta)/\alpha\ , \qquad [12b]$$

$$\theta \rightarrow (\theta + \beta)/\alpha\ . \qquad [12c]$$